Report 78-20
Stanford -- KSL

Scientific DataLink

Use of a Computer to Identify Unknown Compounds: The Automation of Scientific Inference. Neil A.B. Gray, Dennis H. Smith, Tomas H. Varkony, et al., Oct 1978

HPP 78-20

Use of a Computer to Identify Unknown Compounds: The Automation of Scientific Inference.

N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart and B.G. Buchanan.
Departments of Computer Science, Chemistry and Genetics,
Stanford University,
Stanford,
California 94305.

October 14, 1978

Contents.

Table of Contents

BLANK PAGE

Introduction.

A Introduction.

In the period since the publication of the original article, research within the theme of the title has evolved and broadened in scope considerably. In this supplement, the major developments will be outlined, using the previous description of the Heuristic DENDRAL program as a point of departure. The emphasis will be on those new programs which are involved, directly or indirectly, in the utilization of mass spectral data in structural analysis. The reasons for concentrating on particular aspects of the use of computers in structure elucidation will be presented as will the strengths and limitations of the various programs that have been developed.

The principal objective of the DENDRAL project has evolved toward developing an integrated approach to computer-assisted elucidation of molecular and biomolecular structures. The structural problems of interest are those for which X-ray techniques are inappropriate (insufficient or uncrystallizable samples and mixtures), and which must instead be solved by physical, chemical and spectroscopic means. Frequently, mass spectrometry plays an important role in such problems. In the process of developing general techniques to aid the structural chemist, several new programs have been devised which can in some way exploit mass spectral data. These programs include the PLANNER program for inferring structural details of a compound of known class from its mass spectrum (19), the CONGEN program for identifying all possible structures compatible with structural constraints that a chemist has inferred from physical and chemical data (20) and the Meta-DENDRAL programs for developing new rules for the interpretation of mass spectra (21,22).

Considering the general problem of structure elucidation, it is relatively simple to outline the important milestones in the process. The first is the isolation of the compound and the determination of some of its chemical and spectral properties. These initial data lead to various structural hypotheses. Thus, certain types of functionality may be eliminated by considering the source of the compound and nature of the isolation procedure; other functionalities may be identifiable from the spectral data. As well as identifying the presence or absence of specific substructures, it is usually possible to derive additional general constraints on the ways in which the substructures can be assembled. Constrained assembly of identified substructures and residual atoms yields a set of candidate structures for the unknown compound. Examination of these candidates will suggest additional spectral and chemical experiments that might serve to differentiate amongst them.

We have sought ways to enhance the process of structure elucidation by emulating certain parts of the task in computer programs designed to assist chemists in solving structures. We have concentrated

on taking advantage of the potential synergism of man and machine by directing computer developments first to the parts of the problem most difficult to perform manually, e.g. exhaustive and irredundant structure generation. Thus, the chemist is able to devote attention to the interpretive aspects of the problem which are beyond the scope of current computer techniques. Most importantly, we have developed ways to take advantage of structural information from many sources besides mass spectrometry to solve the majority of new structures where the mass spectrum alone is insufficient to establish identity.

B Motivation.

The motivations presented in the earlier chapter still hold true including exploration of methods for manipulating representations of organic molecules in a computer, for generating all possible structural hypotheses for an unknown and for systematizing the generator process so that it can make used of information about structures as heuristics to restrict the generator procedure. However, we no longer restrict out attention to use of mass spectral data exclusively in suggesting plausible structural alternatives as was done in the original Heuristic DENDRAL program. The new programs described in subsequent sections evolved from the original in ways governed by several factors. Although performance was quite high for saturated aliphatic monofunctional compounds, such compounds do not represent a significant subset of biologically important compounds. Before the Heuristic DENDRAL method could be used for molecules of greater complexity, several problems had to be addressed:

1) No method existed for efficient generation of cyclic molecules;
2) No analog existed of the "preliminary inference maker" for determining the compound class and invoking relevant fragmentation rules.
3) The definition of a compound "class", based on molecular topology, becomes less useful in complex polyfunctional molecules where structurally related molecules may fragment along very different pathways.
4) No analog existed of the "predictor" for ranking structural candidates for complex molecules.
5) Complementary structural data derived from techniques other than mass spectrometry were being underexploited.

Because of these factors, development of new computer techniques proceeded along separate lines. One involved development of the PLANNER program which accepted as input a chemist's definition of compound class and fragmentation rules and analyzed mass spectra on the assumption that each spectrum was of a compound or mixture of compounds in that class. This approach thus replaced the preliminary inference

maker with the chemist who could bring a wealth of additional knowledge about the sample to bear on class definitions and relevant mass spectrometry rules. As an aid to determining rules for new classes of compounds the Meta-DENDRAL programs were written (21,22). As a separate effort, a structure generator for cyclic molecules was devised, programmed, and provided with mechanisms for implementing constraints. More recently, the task of ranking candidate structures based on predicted mass spectra has been explored and new results are presented in subsequent sections.

These efforts in principle complete the Plan-Generate-Test paradigm for problem solving discussed in more detail in the next section. However, the programs have not been integrated into a complete system to yield a more general Heuristic DENDRAL program for at least two reasons. The stand-alone programs have proven useful for assisting in solving structures. Mass spectrometry is insufficient for solving many of the structures we have encountered in our work, which is an effective deterrent to building a system which is heavily dependent on mass spectral data alone.

C Implementation.

.1 The Plan, Generate, and Test Paradigm for problem solving.

The DENDRAL systems derive from the PLAN-GENERATE-TEST paradigm for problem solving (23). The kernel of each of the DENDRAL systems is some algorithm for the exhaustive generation of all possible solutions to a problem; these possible solutions would be chemical structures in the case of structure elucidation problems, or rules for predicting spectral features in the rule generation systems. The generating algorithms are designed to exploit prospective constraints that have been inferred by preliminary planning functions of the systems. Typically, the preliminary planning phase will exploit rules of proof and disproof and data that can be related to explicit subparts of possible solutions. Possible solutions are passed to some final testing phase wherein some candidates are eliminated and the remainder are ranked according to some measure of their plausibility. The testing phase will utilize rules that express estimates of plausibilities and such data as can only be interpreted in terms of a complete structure.

The simplest example of the Plan-Generate-Test paradigm within DENDRAL is the MOLION program for molecular ion identification (24). In MOLION, the planning phase involves the determination of the probable

parity of the molecular weight and the identification of those secondary losses observed in a mass spectrum that might correspond to primary losses from the (possibly absent) molecular ion. The generation phase creates candidate molecular ion masses by adding the masses of the identified neutral losses to the masses of ions observed in the high mass region of the spectrum. In the final test phase, candidates at masses below the highest observed mass (discounting isotope peaks) and candidates showing apparent "bad losses" to observed ions are eliminated. The remaining candidates are ranked according to the strength of supporting evidence from the number and intensity of the neutral-loss/fragment ion pairs used to generate the candidate.

Although limited planning programs utilizing mass spectral data, e.g. PLANNER (19) and MDGGEN (25), have been developed as parts of the DENDRAL project, the planning phase in most structure elucidation problems is left to the chemist. Currently, it is for the chemist to determine, from all available chemical and spectral data, what substructures are present and what constraints must be imposed upon the combinations of these substructures. The substructures and constraints form the input of the CONGEN program. CONGEN will exhaustively and irredundantly generate all structures compatible with the given constraints.

Functions have been developed for testing CONGEN generated structures for compatibility with mass spectral data. Possible structures are ranked by functions that estimate how readily the observed spectrum can be rationalized in terms of the allowed fragmentations of each proposed structure; the user can control the types of fragmentation process that the program is to consider. Future developments may include other ranking functions based on magnetic resonance or other spectral techniques. However, there are problems in the exploitation of these other techniques relating to the need to consider stereochemical, geometrical, and conformational influences on spectra, for the computer representation used for structures defines just their topology (with some recent extensions to include stereochemistry (26,27)).

In the "meta-DENDRAL" theory formation programs (INTSUM, RULEGEN, and RULEMOD), the entities being manipulated are rules of the form:

substructural feature ——> predicted spectral process.

The predicted process is a definition of which bonds break in the specified substructure, where charge resides, and what atoms may be transferred between fragments. The planning phase for mass spectral rule formation is the INTSUM program which interprets spectra of known structures, sharing some common skeleton, to produce evidence for processes involving breaks of that skeleton and atom migrations (21).

Implementation.

The generator, RULEGEN, creates rules by selecting features of the environment of fragmentation sites identified by INTSUM; the generated rules are finally filtered by the RULEMOD program (22).

The subsequent sections are organized to correspond with the Plan-Generate-Test methodology. First, in the section on planning, programs that attempt to infer structural information from mass spectral data are discussed. Then, the new algorithms for structure generation are described along with the CONGEN program which provides both a user interface to the structure generator and a means of controlling it by specifying structural constraints inferred by conventional manual interpretation of spectral and chemical data. The section on the testing phase describes approaches to ranking candidate structures based on agreement of predicted and observed mass spectral features. The Meta-DENDRAL programs are discussed in a separate subsection.

.2 The use of mass spectral data in the Planning phase.

Early publications on the HEURISTIC-DENDRAL program emphasized the use of mass spectral interpretation procedures in the planning phase of the structure elucidation process. The PRELIMINARY-INFERENCE-MAKER module of HEURISTIC-DENDRAL used simple hierarchic, classification networks to derive constraints for the structure generator. If a recorded mass spectrum satisfied all the rules characterizing some class in the hierarchy, then the corresponding substructure was entered on the generator's GOODLIST of allowed features; otherwise, the substructure would be added to the BADLIST of prohibited features. Rule networks were created for a number of classes of simple, acyclic, mono-functional molecules including ethers, amines, and ketones (28-31).

Generally, such networks of classification rules cannot be combined to process polyfunctional molecules. One constituent group in a molecule may direct fragmentation so strongly that cleavages about other functionalities are not observed; thus, in polyfunctional structures, the absence of expected ions does not necessarily constitute evidence against the existence of given groups. Since classification rules cannot just be combined, it is necessary to create spectrum interpretation schemes for each new class of structures.

The effort of developing elaborate, class-specific schemes can be justified in applications where many closely related structures need analysis. If, in a given class of compounds, the structures differ only in the nature and position of substituents about some common skeletal nucleus, and if the major fragmentation processes that occur are those cleaving the skeleton, then rules defining the fragmentation processes

of the skeleton can be used to identify the structures. The mass spectral PLANNER program was developed for this, somewhat limited, class of structure elucidation problems (19).

In contrast to the class specific-methods of the HEURISTIC DENDRAL and PLANNER programs is a new general approach embodied in the MDGGEN program for deriving structural constraints from mass spectra (25). This approach is based on the hypothesis that every ion represents an intact portion of the original molecule. Parts corresponding to different ions overlap to unknown extents but candidate structures can be generated by determining all possible overlaps of the pieces represented by different ions. The MDGGEN program is very much an experimental tool. In principle, it can be used to infer constraints for the CONGEN structure generator. But, as yet, there is no automated method for exploiting results from MDGGEN within CONGEN.

.2.1 Class-specific rules and the PLANNER program.

The PLANNER program, for the interpretation of high resolution mass spectra of compounds with known skeletons, is a complete system with its own structure generator and testing functions (19). PLANNER was created before the general purpose structure was available; its own specialized structure generator is limited to producing all differently substituted variants of a common, user-defined, skeleton. Its test phase is also very limited; it provides only for those structural constraints that could not be employed by the very simple structure generator. The spectral data input to the program consists of an unknown's high resolution mass spectrum with ion compositions and intensities, metastable data and, if available, the low ionizing voltage spectrum for helping to identify the molecular ion. The class-specific information consists of a definition of the molecular skeleton and the processes by which that skeleton fragments. The process definitions detail which bonds break, where charge resides, and what hydrogen transfers are permitted.

There are three distinct phases in PLANNER corresponding to Plan, Generate, and Test. The ANALYSIS (Plan) phase involves firstly the characterization of the molecular ion, and consequent determination of the number and type of substituent atoms, and secondly, by use of the given fragmentation rules, the identification of all placings of the substituent atoms for which evidence exists in the spectrum. The SYNTHESIS (Generate) phase combines evidence from the separate fragmentation processes to yield possible substituent placings consistent with all processes. The structures generated in the SYNTHESIS phase are finally checked in the FILTER (Test) routines where any additional constraints, e.g. knowledge of the number of constituent hydroxy groups, can be exploited.

Implementation.

The most important part of the PLANNER program is the set of functions that analyze the spectrum, in terms of the fragmentations of the known skeleton, to determine possible placings of substituent atoms. The operations of these functions can be illustrated through an example of their processing of substituted estrogenic steroids. The estrogen skeleton, and the break processes used by the program are shown in Figure 7-6.

FIGURE 7-6 (BREAK PROCESSES ON ESTROGENIC SKELETON) about here.

The basic nucleus has a composition $C_{18}H_{24}$. The principal break processes, with the expected ions, are: B: $C_{15}H_{18}$; C: $C_{12}H_{13}$; D: $C_{13}H_{15}$; E: $C_{11}H_{12}$; F: $C_{10}H_{10}$; (all possibly modified by the transfer of one or two hydrogen atoms into or out of the charged fragment).

The compound $C_{18}H_{22}O_2$, with the low resolution spectrum shown in Figure 7-7, has two oxygens and an unsaturation to be placed upon the standard skeletal nucleus. For convenience, PLANNER treats the unsaturation as two "DOT" substituent atoms. By considering the possible placements about the sites of cleavage, a variety of different compositions can be predicted for the corresponding fragments. Thus, ions corresponding to process B could appear at $C_{15}H_{18}$ (all substituent atoms attached to skeletal atoms C-15, C-16 or C-17), $C_{15}H_{18}O$ (one oxygen on the charged fragment), $C_{15}H_{18}O_2$ etc. Frequently, evidence will be found for more than one possible distribution of substituents; in such cases, intensity data provide a measure of the relative plausibility of the different possibilities. In this case, examination of the spectral data yielded the possible substituent placings detailed in Table 7-III.

break	substituents on charged fragment	relative weight
B	O_1	
C	(O_1	23
	($O_1 DOT_2$	9
D	O_1	
E	(O_1	25
	($O_1 DOT_2$	16
F	O_1	

Table 7-III.
Conclusions about substituent placings from PLANNER's ANALYSIS phase for compound (1).

FIGURE 7-7 (SPECTRUM OF ESTRONE) about here.

These possible substituent placings are passed to the SYNTHESIS routine that determines those which are consistent with all of the breaks. The procedures can be illustrated for this simple example (just considering the more plausible substituent placings suggested by breaks C and E). Process B leads one to infer that there is one oxygen attached to one of the atom set (C-1 to C-14 or C-18) with the other oxygen and "DOT" substituents on atoms (C-15, C-16 or C-17), while break C places an oxygen on (C-1 to C-10, C-14 or C-15) and the other oxygen and DOTs on (C-11, C-12, C-13, C-16, C-17 or C-18). Taken together, these breaks localize one oxygen and the DOT substituents to atoms C-16 or C-17 of the nucleus and the other oxygen on (C-1 to C-10 or C-14). Break D further localizes the oxygen and both DOTs to atom C-17 (i.e. a carbonyl group at C-17) while processes E and F limit the remaining oxygen substituent to somewhere on atoms C-1 to C-10. As the program had no rules for fragmenting the A or B rings of this skeleton, this oxygen could not be further localized; but, within these limits, the program correctly achieved the identification of estrone (1).

STRUCTURE 1 (STRUCTURE OF ESTRONE) ABOUT HERE.

Implementation.

PLANNER has been used for the analysis of estrogenic steroids; the program's capacity for identifying components in unresolved mixtures has also been demonstrated (32). Difficulties in structure identification occur mainly for those compounds where the observed spectrum is dominated by fragmentations of substituent groups, as in benzoate derivatives of the estrogens, or where substituents induce major changes in the fragmentation behavior of the nucleus.

PLANNER: Strengths and limitations.

The strengths of PLANNER lie in its systematic considerations of the mass spectral evidence and corresponding structural possibilities, and in its potential generality for application to other unknown compounds in a known class. Its limitations are that few structural problems fall neatly within the program's competence. If the class and fragmentation rules are well known, then frequently conventional, manual, interpretation can provide the solution. A further limitation for persons outside of Stanford lies in the difficulty of rapid transfer of large quantities of mass spectral data to the SUMEX facility (see EXAMPLE section). Recent developments of CONGEN and of more sophisticated mass spectral testing functions have largely taken over the role of problem solving previously handled by PLANNER.

.2.2 Mass Distribution Graphs and the MDGGEN program.

The MDGGEN program constructs "mass distribution graphs" suggested by the ions observed in a compound's mass spectrum. A Mass Distribution Graph ("MDG") is a graph whose nodes represent groups of atoms and whose edges represent the existence of bonds between atoms assigned to different groups. These graphs, which when fully elaborated represent complete structures, are created under the guidance of a "theory" of allowed fragmentation processes that could give rise to the observed ions. MDGGEN requires fragmentation rules that specify values for control parameters such as limits on the number of bonds that may be broken and details of allowed hydrogen transfers and neutral losses. The operation of MDGGEN can be illustrated through its processing of mass spectral data of hexanal (structure 2). The low-resolution mass spectrum of hexanal is shown in Figure 7-8. Compositions were determined for the fragment ions by accurate mass measurements from HRMS. The eight fragment ion compositions shown in Figure 7-8 form the input data data for the MDGGEN program.

FIGURE 7-8 (SPECTRUM OF HEXANAL) about here

STRUCTURE 2 (STRUCTURE OF HEXANAL) about here

The initial MDG is just a single node representing the molecular composition. The program expands existing MDGs into more elaborate, detailed MDGs that can serve to rationalize the appearance of additional ions.

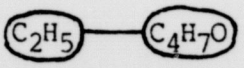
The program operates by selecting individual ions and attempting to apportion atoms in nodes of the current MDG amongst nodes of new MDGs such that some allowed fragmentation of these new MDGs would yield the selected ion. Beginning with the molecular ion MDG ($C_6H_{12}O$) and using the observed ion C_3H_7 results in the new possible MDGs shown in Figure 7-9. Simple bond cleavage of the first MDG would yield C_3H_7 directly; if hydrogen transfers are allowed then the second MDG is also possible. If two-step cleavages are permitted, then five additional MDGs may be generated all of which could yield C_3H_7 with the indicated cleavages and hydrogen transfers.

FIGURE 7-9 (MDGS ETC) about here.

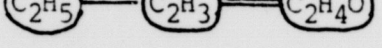
The process of expanding MDGs is controlled by a function that determines the allowed overlaps of existing MDGs and new graphs implied by the next ion. For example, MDGs representing possible ways of generating the ions C_4H_8 and C_2H_5 from this molecular composition are:

(g1) 

and

(g2) 

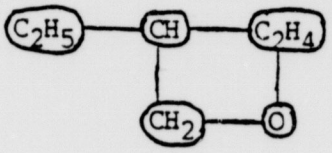
The only possible overlap of these yields the more specific graph:

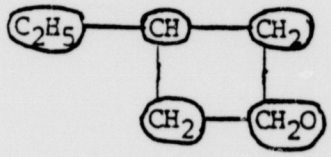
(g3) 

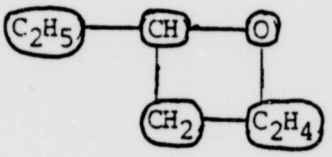
More typically, several alternative more detailed graphs result from performing such an overlap operation. Overlapping of graph (g3) with graph (g4)

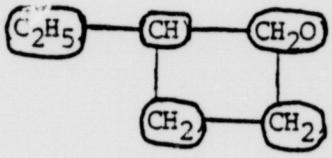
(g4) 

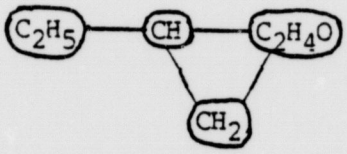
which represents one of the possible ways of obtaining the ion C_3H_7, yields five possible MDGs (g5–g9). Each of these more specific MDGs provides some basis for rationalizing the ions C_2H_5, C_3H_7 and C_4H_8.

(g5) 

(g6) 

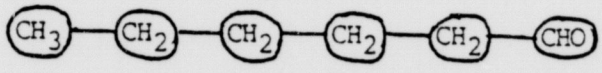
(g7) 

(g8) 

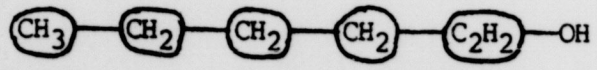
(g9) 

The generating procedure utilizes valence and parity constraints that prevent the production of meaningless MDGs and, in addition, employs the constraints provided by the user-defined fragmentation rules. If, for example, the fragmentation rules did not permit breaks of adjacent bonds on the same carbon, then MDGs (g5), (g6), (g8) and (g9) could all be discarded.

The fragmentation rules used to process the data on hexanal permitted only single-step processes with no cleavages of multiple or adjacent bonds, allowed transfers of 0 or 1 H-atoms and required that not more than one ion go unexplained. Under these constraints, only two MDGs resulted. The first (g10) corresponds to hexanal while the second (g11) represents two isomeric enols.

(g10) 

Implementation.

(g11) 

MDG (g10) accounts for all ions save C_6H_{10}, while (g11) fails only to explain the CHO ion. Cyclic graphs, such as g7, though generated at intermediate stages in the computations were all eliminated for they could not account for adequate numbers of observed ions.

MDGGEN: Strengths and limitations.

The strength of this approach lies primarily in its exhaustive consideration of structural possibilities independent of the compound class or class-specific rules. However, the inherent combinatorial complexity of the task limits the utility of MDGGEN. Even with severe constraints, there are usually several MDGs resulting from the application of a single ion to each of the MDGs resulting from previous steps. The task of determining all overlaps for several ions in a spectrum of a molecule of significant size is time-consuming, in part because of the current lack of a means for exploiting chemical constraints. Further, there is considerable duplication inherent in the procedure. Observed ions may frequently be explained in terms of many different MDGs. However, these different MDGs may all yield the same structure because they correspond to alternative ways of fragmenting that structure to produce the ions. Another limitation is that while the fragmentation theory used must be restricted to constrain the generation procedure, excessive restrictions on the permitted fragmentations will result in ions that cannot be explained and possibly a consequent failure to generate the correct structure. Thus, MDGGEN remains an interesting research project but it is not yet a useful program for structural studies.

.3 CONGEN: the constrained generation of chemical structures.

The DENDRAL project originated in the work by Lederberg on the enumeration of organic molecules as tree structures and cyclic graphs (33). The number of chemical graphs for any molecular composition of interest is extremely large (34,35); it is this combinatorial complexity that would defeat any approach to structure identification that relied solely on the exhaustive generation of possible candidates and subsequent testing for their compatibility with available data. Research in the DENDRAL project has been aimed at finding ways to exploit information at the earliest possible stage in order to most effectively constrain subsequent generation steps. The HEURISTIC DENDRAL program established the efficacy of employing heuristics (empirical rules for inferring constraints from data) to guide a

structure generator. Because of complexities relating the generation of cyclic structures, this initial work was limited to simple, acyclic molecules (28-31).

A structure generator was subsequently developed that could produce an exhaustive, irredundant set of structures with arbitrarily complex cyclic and acyclic components (36,37). This structure generator breaks the problem down into many separate subproblems by determining all possible ways of dividing up the available atoms and unsaturations into one or more cyclic components and acyclic portions. Each distinct way of partitioning the atoms yields a separably soluble structure generation problem. Within a given partition, each cyclic component can also be treated as a separate subproblem.

The algorithm for processing individual cyclic components is based on the concept of "vertex-graphs". Vertex-graphs were originally used by Lederberg for classifying cyclic structures (33); a vertex-graph consists of vertices (nodes) which represent the points of ring fusions in purely cyclic structures together with the edges that interconnect these nodes. Figure 7-2, in the earlier version of this chapter, illustrates several graphs constructed with just trivalent vertices. Vertex-graphs can be indexed according to the number of nodes of each given degree that they possess; early tabulations of vertex-graphs of degree three and four have been greatly extended as a part of the development of the cyclic generator (38).

The underlying vertex-graphs appropriate to a particular cyclic component are determined by functions of the number of atoms of each degree and the number of unsaturations assigned to that component. These standard vertex-graphs are expanded through a sequence of steps to yield complete ring-systems containing nodes representing all constituent atoms of the cyclic component and having each node tagged with atomic symbol and any necessary free valences (which represent points of attachment to acyclic components or other ring-superatoms).

Once all the cyclic components have been processed to give sets of ring-superatoms, final structures for a particular initial partitioning of the atoms can be derived by assembling all alternative selections of ring-systems and those atoms originally assigned to the acyclic part. The functions used for assembly are similar to the acyclic structure generator described by Lederberg in the earlier version of this chapter.

As demonstrated in the HEURISTIC DENDRAL exercises, if one can identify substructures containing many of the constituent atoms of a molecule, then great reductions can be made in the subsequent generation procedures by using just the residual atoms and "superatoms" representing the known substructures. The complete structure generator, and functions for using predefined superatoms, underlie the CONGEN

isomer generation program (20). The objective of CONGEN is the CONstrained GENeration of all isomers of a given molecular composition. By means of an interactive executive routine, CONGEN allows a chemist to specify the molecular composition, to define superatom parts representing known substructures, to express additional constraints upon the combinations of substructures and residual atoms and to control the operation of the generator and imbedder (20,39). CONGEN also includes an interactive structure editor (EDITSTRUC), a teletype-oriented structure display program (DRAW) (40), functions for surveying sets of generated structures to determine the frequency of standard functional groups and skeletons (EXAMINE), and file manipulation utilities for saving and restoring problems.

Applications of CONGEN, by users in Stanford and elsewhere, have included structure determination of terpenoids (41), steroids (42), organic acids and other constituents of body fluids (25), metabolites in micro-organisms, and photochemical rearrangement products. Use of CONGEN is illustrated in the EXAMPLE section of this article.

CONGEN: Strengths and limitations.

CONGEN's strength is its ability to generate, exhaustively and irredundantly, all isomers compatible with given structural constraints. CONGEN's generator, and superatom manipulating, functions are based upon mathematically proven algorithms (43-45). However, by deliberate choice, CONGEN works just in terms of graph theory, with recent extensions to configurational stereochemistry (26,27), and has no internal model of chemical stability; consequently, unless prevented by user-defined constraints, CONGEN will produce many totally unacceptable isomers involving highly strained ring systems or incredible functionalities. While CONGEN allows the user to define a very broad range of structural constraints inferred from arbitrary chemical and spectral data, it cannot assist the user in interpreting the available data to yield such constraints.

.4 The analysis of mass spectral data in the Test phase: MSPRUNE, MSRANK and related functions.

A variety of models, or "theories", can be used for predicting the ions that could arise by the fragmentation of a molecular skeleton. The type of fragmentation theory to be used depends largely on the context of the structure determination problem. When initially studying a new class of compounds, or when attempting to discriminate between different candidate structures obtained from some unusual CONGEN

problem, it is usually appropriate to use some universal form of fragmentation theory that expresses just general chemical principles. One can be confident that a general theory will apply to the structures and will not be biased in its predictions. However, a general theory may well prove to have poor discriminatory power. Fine discrimination between related structures generally requires more refined fragmentation theories wherein one assigns different plausibilities to alternative fragmentation processes. When processing isomers from some well characterized class, the appropriate fragmentation theory may well involve the detailed specification of substructures, their bond cleavage processes, and the accompanying specific transfers of hydrogen atoms or other molecular fragments. We have developed a single set of functions that can accommodate theories of differing degrees of generality. We shall illustrate the application of these theories to the analysis of a variety of classes of chemical compounds studied in our laboratory. First, we show how even the most general fragmentation theory, "half-order theory", can serve to discriminate between isomers of moderately complex structures such as monoketoandrostanes. Refinements of the simplest half-order theory involve first the use of estimates of relative plausibilities of fragmentation processes of differing degrees of complexity and, subsequently, the association of relative plausibility values with vinylic, allylic, and other classes of bond cleavage. These refinements are illustrated through an example analysis of steroids. Finally, the use of detailed class-specific fragmentation processes is considered in relation to the processing of the spectra of macrolide antibiotics.

.4.1 The Half-Order Theory of molecular fragmentation.

The simplest model of molecular fragmentation is the ALLBREAKS "theory" that allows for all ions formed by any combination of bond cleavages and transfers of atoms between fragments. Such a model is too general and has no application in computer analysis of mass spectra. However, a constrained version of the ALLBREAKS theory can constitute a usable method of spectrum prediction. A very simple version of a constrained ALLBREAKS approach was used in the Heuristic DENDRAL program described in the earlier version of this chapter; this early version was designed just for acyclic, monofunctional molecules (46). DENDRAL's "half-order theory" of mass spectrometry is a more general example of such a constrained ALLBREAKS model of molecular fragmentation. The half-order theory is a very loose model; it does not describe the detailed aspects of fragmentations, such as specific relationships between atom transfers and cleavage of certain groups of bonds, and it makes no attempt to express anything about the mechanisms by which fragmentations actually take place.

The constraints that can be expressed in the half-order theory of molecular fragmentation include limitations on the number of bonds

that may be broken, the number of steps in a process, the proximity of pairs of breaks (i.e. whether or not two adjacent bonds can break in a given process), the multiplicity or aromaticity of each cleaved bond, the allowed hydrogen transfers into or out of the charged fragment, and the neutral fragments that may be lost. Functions have been written that, given a set of constraints and a structure, will predict the corresponding spectrum. Each predicted ion is formed by a "process" involving

i) one or more cleavage "steps".
ii) possible H-transfers.
iii) possible neutral losses.

Each "step" cleaves a molecule and may be a break of one acyclic bond, two bonds within a ring, or a group of three bonds in a fused or bridged ring system. The most elaborate possible step is such a fused/bridged ring cleavage as shown in scheme 7-D.

---

scheme 7-D about here.

A complete process could for example involve fused-ring cleavage, simple ring cleavage, and acyclic bond cleavage steps; such a process would involve a total of six bond breaks. Typical constraints used with the half-order theory would be:

i) prohibit cleavage of aromatic or isolated double or triple bonds,
ii) allow one or two step processes.
iii) allow at most two bonds to be cleaved in a given step.
iv) permit a maximum of three bonds to be cleaved in a process.
v) prohibit the cleavage of two (non-hydrogen) bonds from the same carbon atom, (for this would formally leave a charged carbene species which is normally energetically unfavorable).
vi) restrict transfers between fragments to at most two hydrogen atoms.

Even in this very limited form, the half-order theory can be of value in helping to identify candidate structures compatible with spectral data. For example, both of the monoketoandrostanes (structures 3 and 4) could conceivably fragment to give the ion $C_8H_{10}O$:

STRUCTURES 3 AND 4 ABOUT HERE.

However, in structure (3) only a simple ring cleavage and hydrogen transfer are required whereas for (4) to yield this ion would necessitate cleavage of a fused-ring system. The MSPRUNE function uses such differences to eliminate candidate structures (25). MSPRUNE takes as input a set of CONGEN-generated structures, parameter values for the half-order theory, and ion compositions of those ions that the user requires to be explicable in terms of a given structure under the specified fragmentation constraints. If, in this example, the user limited the half-order theory to a maximum of two bonds broken in any step, then structure (4) would be rejected.

Generally, we have found it to be more effective to employ the data in the entire observed mass spectrum, and rank candidate structures according to how well they serve to explain the spectral data. This ranking is accomplished through the MSRANK function which allows the user to define the constraints of the half-order theory and to specify the form of the scoring function. The score assigned to a candidate structure is determined from the importance accorded to those of the observed ions that can be generated by the allowed fragmentations of that candidate. As ions at higher mass and intensity values are generally of greater structural significance, the importance accorded to each observed ion in the spectrum is determined by some function of its m/z and its relative intensity (in most cases, the product of m/z and intensity has been used).

The results shown in Table 7-IV typify the performance of this simple approach to discriminating between structures. The structures analyzed were monoketoandrostanes, all with the same steroid skeleton but varying in the position of the keto substituent. Using the simple half-order theory approach, the eleven possible isomers were ranked against each of the ten available high resolution mass spectra. The half-order theory generally separates structures into two groups, those that match the spectra about equally well and those that can definitely be eliminated. With these structures, the correct candidate was generally ranked first when comparing its predicted and recorded spectra but it was not possible to discriminate between different isomers with the keto group on ring A or amongst those with the keto group on ring D.

STRUCTURE (KETO POSITION)	RANKING	STRUCTURES WITH EQUAL SCORE	BETTER RANKED STRUCTURES
7	2		6
16	1	17,15	
11	2		12
3	1	1,2,4	
17	2	15,16	1,2,3,4
6	1		
12	1		
15	1	17,16	
1	1	2,3,4	
4	1	1,2,3	

Fragmentation constraints:

one-step processes, a maximum of two bonds cleaved, transfer of at most two hydrogens into or out of the charged fragment.

Table 7-IV.
Ranking of monoketoandrostanes based on the half-order theory.

.4.2 Half-order theory with process and bond break plausibilities.

In cases where the differences between candidate structures are not simply related to the major fragmentation processes, then generally little discrimination can be achieved by MSRANK's just testing for the existence of fragmentation processes that could account for observed ions. However, in such cases, we can use the MSRANK program with a more refined version of the half-order theory in which relative plausibility values, in the range 0-1, are associated with processes of involving differing numbers of steps, differing numbers of bond cleavages and different types of neutral transfers between fragments. The principle behind this is that if two structures both provide explanations for an observed ion, then the structure with the simpler explanation is more likely. The plausibility of a predicted ions is given as the product

of the break plausibilities of the bonds and H-transfers or neutral losses involved, modified by any additional factors such as the reduced plausibility of process requiring adjacent breaks or multiple steps. If an observed ion can be rationalized in terms of two different fragmentations of the same structures, then the process with the higher plausibility is used. The scores assigned to a structure take the sum of the product of the process plausibility and the importance (m/z times intensity) of the ion for each observed ion that can be generated. With these plausibility weighting factors, the half-order theory can discriminate quite well between related structures. Typically, around half or even two-thirds of a set of candidate isomers can be rejected on the results of mass spectral ranking.

Even with very closely related structures, some degree of discrimination can usually be achieved. A typical example involved the processing of sterols sharing the same steroidal skeleton and differing only in the form of their C_{10} olefinic side chains. The isomers varied in the position of a double bond in, and branching of the side chains. These structures were amongst those generated by the REACT program in a study of the biosynthesis of natural products (47); all could be derived from an initial sterol with a C_8 side chain by known biosynthetic reaction pathways. These structures were selected to assess how well a mass spectrum could characterize the correct member of a set of really closely related isomers; since these structures differ just in respect to a feature that does not strongly direct fragmentation, only very minor differences are expected in their spectra. All were isomers of fucosterol (5).

STRUCTURE 5 (STRUCTURE OF FUCOSTEROL) about here.

Fucosterol, and 39 of its isomers, were ranked against the recorded high resolution mass spectrum of fucosterol. Previous studies of the mass spectral behavior of such steroids had established general guidelines that could be used to estimate plausibility parameters for the half order theory. Thus, we knew that most ions in these spectra can be attributed to one-step processes, usually involving no more than two bond breaks. We also knew that most fragment ions were accompanied by others corresponding to further loss of 15 or 18 u and that transfers of hydrogens out of the charged fragment were more common than hydrogen transfer into the charged fragment. The fragmentation control parameters given in Table 7-V express this general knowledge about this class of steroid fragmentation:

Constraint				Value
maximum number of steps				1
maximum number of bonds broken				2
plausibility of process with one bond break				1
with two bonds breaking				0.85
plausibility of adjacent breaks				0.25
plausibility of breaks of multiple bonds				0
plausibility of loss of CH_3			1.	
plausibility of loss of H_2O			0.85	
Hydrogen transfers	-2	-1	0	1
and plausibilities	1.0	1.0	1.0	0.85

Table 7-V.
Fragmentation Process Parameters used in the analysis of the Fucosterol (5) data.

With these parameters, MSRANK was able to eliminate six of the candidates, mainly those with a C-22 - C-23 double bond or those with an alkyl substituent on C-22. The structures eliminated all had scores of 30% or less; further discrimination between the remaining thirty-four candidates was not possible as all had scores in the range 55% to 60%.

Using additional knowledge about the relative stability of bonds in these structures, we can define relative plausibilities for specific bond breaks. Thus, the relative plausibility of allylic cleavages can be increased while those of vinylic bonds, or bonds between secondary carbons in an alkyl chain, can be reduced. Use of substructural features can substantially improve the discriminatory power of the ranking scheme. In the case of the isomeric fucosterols, the general half-order theory could make no significant distinction between structures with the different forms of side chain shown in Figure 7-10.

FIGURE 7-10 (two side chains) about here.

However, in one case a class of important ions arising from the C-22 - C-23 break would require an unfavorable vinylic cleavage, while in the other they would arise through a favored allylic cleavage.

Implementation.

The data on the sterols was reanalyzed, using a more complete theory to define bond break plausibilities. The new parameters are summarized in Table 7-VI. With these constraints, at least fifteen and possibly as many as twenty-two of the forty isomers could be eliminated. Fucosterol was one of just eight, equally ranked, best candidates.

SUBSTRUCTURAL FEATURES: substructure	cleavage	relative plausibility.
C-C=C (*)	vinylic	0.3
C-C-C=C (*)	allylic	1.0
C(C)-C(*)-C(C)	alpha to tertiary & quaternary carbons	1.0
C(C)-C(*)-C-C	alpha to secondary and quaternary carbons	0.9
other single bond breaks, plausibility:		0.8

Table 7-VI.
Bond break plausibilities defined in terms of substructures for use in the more detailed analysis of the Fucosterol data.

.4.3 The use of class-specific fragmentation rules in MSRANK.

Even with refinements, such as the use of substructures to define the relative plausibilities of different bond breaks, the half-order theory typically fails to discriminate within a large group of equally well ranked structures. In part, the degeneracy of scores is

inevitable for the candidates will have very similar structures. However, the over-generality of the theory also contributes to this degeneracy of scores. Half-order theory does not allow specific hydrogen transfers or neutral losses to be associated with specific break processes. Either such transfers are ignored, resulting in explanations being generated for only a fraction of ions in the spectrum or, these transfers are permitted for all possible molecular fragmentations, resulting in ions explained in terms of combinations of fragmentations and transfers which do not in fact occur.

In structures with several charge-localization/fragmentation-directing substituents, interactions between competing fragmentation processes must be expected. It is hard to predict a priori the result of such interactions on the appearance of the mass spectrum and general "half-order theories" can be of limited value. However, if the candidate structures are from a previously studied class, then rules defining their fragmentation behavior can be used instead of, or as a supplement to, half-order theory in the process of generating explanations for observed ions and ranking structures. Generally, the use of fragmentation rules allows for a finer discrimination among candidates.

The rules given to the program must specify substructures, bond breaks and specific transfers. Such rules have been derived in our laboratories for a number of compound classes including the macrolide antibiotics. Some of the standard macrolides are shown in Figure 7-11. Conventional analysis, and results from the INTSUM program, showed that the major fragmentations of the macrolide skeleton could be described in terms of certain McLafferty rearrangements, cleavages alpha to carbonyl groups, and other processes illustrated in Figure 7-12. To test the discriminatory power of the fragmentation rules that had been derived, isomers of the standard macrolactones were generated using CONGEN. In each case, the standard macrolactone skeleton was retained, the isomers varying only in the position of hydroxy, keto, and alkyl substituents and in the position of the double bond in the macrolactone ring. The generated isomers were ranked using the experimental mass spectrum of the standard compound.

Figure 7-11 (5 macrolides) about here.

Using just half-order theory, without any plausibility weightings, the spectrum ranking functions could always eliminate at least 75% of the isomers. However, the correct structure was never uniquely identified and, especially in the cases of methynolide and dihydro-neomethynolide, considerable ambiguity remained.

Implementation.

Structure	Total of isomers	number of isomers with equally good explanations of recorded spectrum	
		Half-order theory	Class-specific rules
6	60	5	2
7	105	28	1
8	105	4	2
9	105	27	5
10	105	10	5

Table 7-VII.
Comparison of performance of MS-ranking functions when using half-order theory and class-specific rules. The structures are the macrolide antibiotics shown in Figure 7-11. Sets of CONGEN-generated isomers were ranked using the recorded high resolution mass spectrum of the standard structures.

Figure 7-12 (break processes) about here.

The rules that had been derived for these skeletons involved all pairwise combinations of the single break processes shown in Figure 7-12. Specific hydrogen transfers were associated with each rule. As shown in Table 7-VII, use of the rules in the spectrum generating process resulted in greatly reduced ambiguity. With rules, methynolide could apparently be identified with certainty and, typically, less than one isomer in twenty was found to be compatible with the spectral data.

.4.4 Applicability

Although capable of handling low-resolution data with just nominal masses, these mass-spectral processing functions have been applied primarily to the analysis of data obtained at high resolving power yielding accurate masses and elemental compositions for all ions. Typically, low resolution spectral data do not provide sufficient information to discriminate between the closely related candidate structures. While the correct structure will normally be amongst the

better ranked candidates, ranking against a low-resolution spectrum rarely results in the candidates being divided into a plausible group and a rejectable group.

.4.5 MSRANK functions: Strengths and limitations.

As demonstrated in the examples summarized above, the MSRANK mass spectral testing functions are of general applicability. These functions have sufficient flexibility to accommodate mass spectral fragmentation theories of differing degrees of specificity as may be required in different types of problems. This approach to ranking of candidate structures, through the analysis of mass spectra, has been tested and proven to perform well for a wide variety of chemical structures. However, when interpreting the results from MSRANK, it must always be realized that the fragmentation theories are just simple approximations and that cases can arise where simplifying assumptions used in the theories may be inappropriate.

.5 The discovery of new class-specific rules: The Meta-DENDRAL System.

.5.1 Data interpretation and summary: the INTSUM program.

Several of the DENDRAL programs share a common need for elaborate, class-specific fragmentation rules. The PLANNER program is totally dependent on such rule schemes. While half-order theory (particularly with extensions relating bond break plausibilities to substructures) can serve quite well to predict approximate spectra, fine discrimination between closely related structures also requires more precise fragmentation rules. The identification of the major fragmentation processes for a class of molecules is a complex and time-consuming task. The Meta-DENDRAL program is designed to aid in the inductive task of finding mass spectrometry rules in a collection of spectra of known molecules. First, possible fragmentation processes have to be hypothesized and then the spectra of all available example compounds must be analyzed to derive supporting evidence for each process. INTSUM, part of the Meta-DENDRAL program, has been devised to assist chemists in this phase of spectral analysis.

INTSUM has the usual PLAN-GENERATE-TEST structure of the DENDRAL programs. Planning is fairly limited, and done by the chemist who defines the common skeleton of the compounds under investigation and specifies constraints on the types of fragmentation process that the program is to generate. Working within the framework of these constraints, INTSUM generates all fragmentations of the skeleton. Each

structure/mass spectrum pair is then interpreted in turn as the program seeks evidence for the hypothesized processes. Evidence for common fragmentation processes is grouped and summary output is provided.

The constraints on allowed fragmentation processes are similar to those used when predicting spectra for CONGEN-generated structures (see "half-order theory" above). The chemist may specify the number of steps in a process, the allowed hydrogen transfers and neutral losses, restrictions on the cleavage of aromatic rings and/or isolated double or triple bonds, and whether or not loss of and/or fragmentations within substituent groups are to be considered. The chemist may also limit the minimum size of fragments that need to be considered; low-mass ions rarely retain sufficient structural information to merit consideration.

Each structure in the training set is defined by identifying the skeletal positions and the nature of its substituent groups. From these data, INTSUM can compute the composition of the ions corresponding to each of the permitted fragmentations of the skeleton. Allowances for hydrogen transfers, and any permitted losses or fragmentations of substituent groups, lead to sets of possible ion compositions for each of the fragmentation processes. In the context of the given structure, detection of any of these computed ions constitutes evidence for the general skeletal fragmentation process. Typically, a particular ion composition may be explicable in terms of several processes and serves as evidence in support of each.

INTSUM summarizes all the evidence for each fragmentation process found in spectra of the training set compounds. Each process for which evidence exists is presented together with a list of those molecules displaying evidence for that process. Cross-references are given to any alternative processes that can also explain the ions used as evidence for a particular process. Processes involving different numbers of hydrogen transfers may be combined in this summary. Fragmentation processes are defined by the breaks of skeletal bonds and transfers involved; further specification is unjustified in the absence of additional data, such as results of deuterium labelling experiments identifying the source of transferred hydrogen atoms.

Published applications of INTSUM include the analysis of the fragmentations of the estrogenic steroids and related equilenins (21), and studies of progesterones (48) and juvenile hormones (49).

INTSUM's analysis of the data on four juvenile hormones, structures 11-14, is typical of these applications. These compounds have potential uses for pest control, and there is a need for methods for detecting them at low concentration. INTSUM was used to identify the characteristic fragmentations of these compounds to permit the design of effective selective ion monitoring experiments. The value of

Implementation.

INTSUM is that it greatly simplifies the task of checking through very large quantities of HRMS data to identify those processes which offer a consistent explantion for fragmentations observed with several different structures. More importantly, it provides a guarantee that, within the specified constraints, all plausible processes yielding observed ions have been considered.

Structures 11-14 here

Although the mass spectra of 11 and 12 (Figure 7-13) do not appear to resemble closely the spectra of the 11-substituted analogs, 13 (Figure 7-14) and 14, INTSUM reveals common diagnostic modes of fragmentation. Firstly, two possible processes are suggested to account for the intense ions at m/z 125 ($C_7H_9O_2$) in 11 and m/z 139 ($C_8H_{11}O_2$) in 12 and the corresponding ions in 13 and 14. These possible processes are summarised in Scheme 7-E.

Figure 7-13 Low resolution spectrum of
ethyl(2E,4E)-3,7,11-trimethyl-2,4-dodecanoate

Figure 7-14
low resolution spectrum
of n-propyl(2E,4E)-11-methoxy-3,7,11-trimethyl ...

Scheme 7-E.

Other evidence suggests that the fragmentation process B is the major contributor, probably involving a cyclization to yield an energetically favorable ion (49).

In the mass spectra of 12-14, ions of m/z 111 are observed which is significant because they retain two oxygen atoms. According to INTSUM, there is only one process, shown schematically as process C, Scheme 7-F, which offers a consistent explanation for these data. This is a two-step process involving, first, cleavage of the 5-6 bond (i.e. process B), followed by loss of the R' group of the ester accompanied by transfer of a hydrogen into the charged fragment.

Scheme 7-F

.5.2 RULEGEN and RULEMOD.

The fragmentation processes identified by INTSUM are specified relative to a complete molecular skeleton. INTSUM does not work in terms of bond environments and will not recognize that different cleavages of a molecular skeleton may in fact represent instances of the same underlying process. In order to identify such underlying processes, Meta-DENDRAL next classifies cleavages according to the substructural environments of the bonds involved, and then seeks unifying generalizations of these many specific bond environments. This process of classification and generalization is accomplished by the RULEGEN program, the generation step of Meta-DENDRAL.

RULEGEN first transforms INTSUM's fragmentation process definitions into the form:

subgraph description —> fragmentation process.

The subgraph descriptions define the connectivity of the structure and atomic attributes (atom type, degree of substitution, and number of hydrogens and pi-electrons) out to a prespecified radius (generally two bonds distant) from each bond cleavage site. Equivalent processes may then be identified, by noting equivalent subgraphs, and the evidence for them combined.

Conceptually, RULEGEN commences its search for fragmentation rules with the most general possible candidate "X*X" where the Xs represent unspecified radicals and the asterisk denotes the bond that cleaves (by convention, the charge is associated with the part to the left of the asterisk). Useful rules are created by stepwise refinement of this initial form. The refinement process involves the expansion of the size of the X groups or the definition of attributes of the atoms in existing X groups. Each attribute type can be considered in turn. Thus, at the first level of refinement of X*X, RULEGEN will consider both rules that specify the types of atoms between which the bond cleaves and rules that are based on the number of hydrogens or pi-electrons on those atoms. Each such candidate rule may be checked against the available data by matching the partially defined bond environment against the specific instances of cleavage environments identified in the INTSUM results. Rules for which evidence exists can become candidates for further refinement by additional subgraph expansion or by the specification of further atomic attributes. This process results in a branching tree of many possible rules; search down a particular branch in this tree continues as long as the more completely specified subordinate rules have greater predictive success than their less specific precursors.

In the last phase of Meta-DENDRAL, the RULEMOD program is used to evaluate and refine the plausible rules created by RULEGEN. When

evaluating possible rules, RULEMOD can detect failures, where predicted ions are not observed, and can exploit such negative evidence to make subgraph descriptions more precise. Frequently, there is substantial overlap between the sets of ions predicted by different RULEGEN rules; some rules may totally subsume others discovered earlier. RULEMOD can rank rules that serve to explain substantially the same sets of ions and can use this ranking to select the more useful rules. Through such processes, RULEMOD can both reduce the number of rules by a factor of five or ten and produce rules that result in fewer incorrect predictions.

The RULEMOD program comprises five distinct subtasks: (a) scoring, (b) merging, (c) refining, (d) generalizing and (e) selecting. Firstly, it selects important rules as identified by its ranking algorithm. Rules are scored by a function that combines the average intensity of their predicted ions and a measure of the extent to which the positive instances of the rule outweigh the negative instances. In this measure of success negative instances, i.e. structures which do not exhibit the hypothesized fragmentation, receive a double weighting as do any positive instances where the process provides the sole explanation of some ions. Ions that can be explained on the basis of highly ranked rules are removed from the sets of evidence for less highly ranked rules. Any rule which consequently loses its entire set of support may be eliminated.

In its second step, RULEMOD attempts to merge any subset of the surviving rules that still serve mainly to explain the same data points. The program attempts to find some generalization of such overlapping rules that still accounts for their evidence and yet does not introduce any additional negative instances.

RULEMOD tries to further refine any rules for which there is negative evidence. This refinement process is similar to the rule search procedures in RULEGEN but allows for more limited exploration of fine changes to the subgraph descriptions.

After a final attempt to create more general rules that could subsume several of the remaining candidates, RULEMOD again ranks its candidate rules. Selected rules from this final ranked set are output.

As well as demonstrations on well characterized classes such as amines and estrogenic steroids, published applications of the INTSUM / RULEGEN / RULEMOD system include the discovery of fragmentation rules for ketoandrostanes (22).

D Commentary.

The programs described above were developed in response to our perceptions as to what, within feasible computational and mathematical resources, would most benefit the structural chemist. To some extent, the programs also reflect the types of structural problems arising in our own laboratories, for improvements and new developments happen most quickly in those areas where a program has been found to be deficient in confrontations with real-world problems.

The programs are weakest in the Planning phase. Essentially, the problem here is one of representing chemical knowledge in a sufficiently structured form for it to be utilized in a computer program. It is atypical for a complex structure elucidation problem to be soluble given just the one kind of data, e.g. a mass spectrum, and data interpretation procedure. Solution of most structure determination problems requires the consideration of the source of the sample and details of the isolation and derivatization procedures as well as the interpretation of complementary spectral data obtained by a variety of techniques. As yet, we have no means for encoding rules for inferring structural details from such diverse data.

CONGEN now offers a complete solution to the problem of identifying all isomers compatible with given structural constraints. The underlying basis of mathematically proven algorithms (43-45), and the thorough testing to which the program has been subjected, together give us confidence in CONGEN's correctness.

The testing phase in structural analysis can use a variety of simple theoretical models for predicting spectra of candidate structures. Thus, for mass spectra we have half-order theory while Shoolery additivity rules could be used to predict proton NMR spectra. Candidate structures can be ranked in those cases where there is some reasonable algorithm for matching the predicted and observed spectra. Predictive models such as half-order theory and Shoolery rules are known to be gross oversimplifications; but some allowance for this can be made when interpreting results obtained thereby. These weak theoretical models are used only to separate candidate structures into a group that appear reasonable in the context of given spectral data and a group which apparently can offer no reasonable explanation of the observations.

The combination of CONGEN and the MSRANK functions does provide a general mechanism for exploiting the knowledge of the chemist and for making some use of mass spectral data in the analysis of candidate structures. Using this approach, the job of initial interpretation of the data is left to the chemist. Through knowledge of the sample, familiarity with related structures and by intuition, the chemist can identify constituent substructures of the unknown and use these in

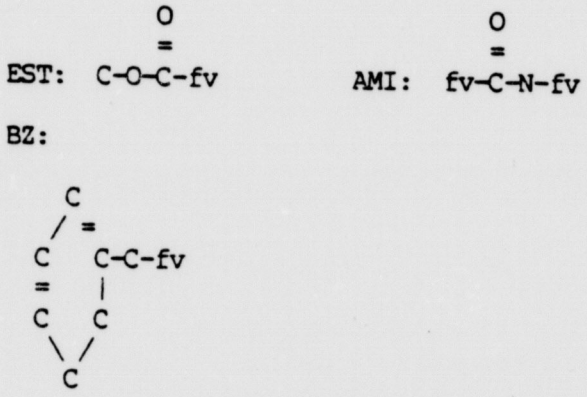
CONGEN. Subsequently, the mass spectrum can be utilized in several ways to rank candidate structures. The ranking process is again left to the discretion of the chemist, who can choose a fragmentation theory appropriate to the particular type of structures being examined. In the example below, we present a simple structural problem which is illustrative of the general approach that can be taken using CONGEN, MSRANK, and related functions for exploring structural possibilities.

E Example. (Diagnosing the Structure of an Unknown Compound arising from an Abnormal Metabolic Pathway).

A simple, but illustrative, application of the CONGEN structure generator and MSRANK structure-ranking functions arose during the course of a study of the metabolic profiles of patients suffering from phenylketonuria. One such patient was placed on an artificial (low phenylalanine) diet; one of the first urine samples investigated after initiation of the diet displayed a large component in the derivatized organic acid fraction which was not present before the diet was begun. Standard analytical procedures, including GC/MS analysis and mass spectral library searching (see Chapter 3), were unable to determine a structure for the unknown.

The molecular formula of the unknown was inferred from the low resolution mass spectra of both methyl and trimethylsilyl (TMS) esters and the GC/HRMS spectrum of the methyl ester as $C_{11}H_{13}NO_3$ (underivatized). (The low-resolution mass spectrum of the TMS derivative is shown in Chapter 3 as an "unknown" compound). Based on the chemical history of the sample and information obtained from the mass spectrum, the partial structures, shown in scheme 7-G, were inferred for the methyl ester (the "fvs" indicate points of attachment to other atoms or superatoms):

Example.



Scheme 7-G

These partial structures were given, together with the molecular composition of the methyl ester, to the CONGEN program. Twelve possible structures were generated; these structures were ranked using the HRMS recorded for the methyl ester and checked for the presence of standard amino-acid skeletons. The following recording of the computer dialog shows the various steps in this process (the recording has been edited slightly to remove irrelevant output, user input is shown underlined):

```
|
|
|    CONGEN 14-Jun-78...
|    DO YOU WANT TO SPECIFY AN EMPIRICAL FORMULA?(Y FOR YES):Y
|    EMPIRICAL FORMULA (? FOR HELP):C 12 H 15 N O 3
|    #EDITSTRUC
|    NAME:BZ
|    STRUCTURE TYPE:S
|    (NEW SUPERATOM)
|    >RING 6
|    >JOIN 1 2 3 4 5 6
|    >BRANCH 6 1
|    >ATOMFV 7 1
|    >DONE
|    (BZ DEFINED)
|    (BZ ADDED TO THE USERATOMS LIST)
```

(The superatom parts for EST and AMI are similarly defined. Then, these known superatom parts are entered on to the composition list and the GENERATE command used to produce structures based on these superatoms and the residual carbons.)

Example.

```
#COMPOSITION

SUPERATOM NAME:BZ
NUMBER OF BZ'S:1

SUPERATOM NAME:EST
NUMBER OF EST'S:1

SUPERATOM NAME:AMI
NUMBER OF AMI'S:1

SUPERATOM NAME:

CHECKING AGAINST THE EMPIRICAL FORMULA...
THE FOLLOWING NON-H ATOMS REMAIN FROM THE EMPIRICAL FORMULA:
TWO C'S
THEY WILL BE ADDED TO THE COMPOSITION LIST
THE RESIDUAL UNSATURATION IS 0 DOUBLE-BOND EQUIVALENTS
#GENERATE
6 STRUCTURES WERE GENERATED
#DRAW 1-2

1
     A   E
BZ-C-M-C-S
     I   T


2

   A     E
BZ-M-C-C-S
   I     T
```

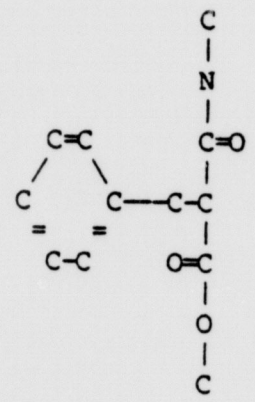
(The structures resulting from this generation step are incomplete for the superatoms are just represented as single atoms. To obtain final structures the superatoms must be expanded out to represent their full internal structure. This is achieved by a call to the IMBED function. Since the superatom AMI can be incorporated in two different ways in each of the partial structures, the number of candidate structures is doubled by this imbedding step.)

```
#IMBED
SUPERATOM NAME:EST
SUPERATOM NAME:BZ
SUPERATOM NAME:AMI
SUPERATOM NAME:
12 STRUCTURES WERE OBTAINED
```

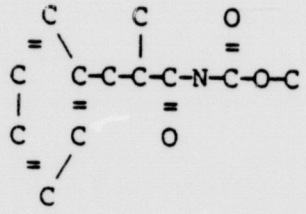
Example.

#DRAW 1-2

1



2



(The twelve structures were passed to the mass spectral analysis functions where the candidate structures were ranked by comparison with the recorded high-resolution mass spectrum. The chemist chooses to specify the parameters for the half-order theory, rather than just accept defaults, for he wishes to allow for breaks of adjacent and multiple bonds (both defaulting to zero plausibility). Otherwise, as substructures are not used to define bond break plausibilities, the parameters are completely general and not specially tuned to these structures.)

#MSA
FILE FOR OBSERVED SPECTRUM: (SMITH)UNKNOWN-METHYLESTER.HRMSSPEC
Parameters: (DEFAULT/FILED/NEW)NEW

Define plausibilities of bond breaks:
single bonds Plausibility:1
aromatic bonds Plausibility:0
double bonds Plausibility:0.2
bond order 3 or more, Plausibility:0

Example.

```
Allow adjacent breaks with Plausibility:0.5
Do you want to define bond-break plausibilities
  in substructures?:N

----------------------------
Molecular ion Plausibility:1

----------------------------
Define complexity of fragmentation processes
Allow fragmentation of fused/bridged rings?:N
Allow simple ring fragmentations?N
ONLY ACYCLIC BONDS WILL BE BROKEN
Max steps per process:2
2 step process, Plausibility:0.5
Max bonds to break in a process:2

----------------------------

Specify possible neutral losses and their plausibilities.
Neutral loss:

----------------------------
Any hydrogen transfers?Y
Possible H transfers: -2 -1 0 1
Enter plausibilities.
Transfer of -2 Hs, Plausibility:0.25
Transfer of -1 Hs, Plausibility:1
Transfer of 0 Hs, Plausibility:1
Transfer of 1 Hs, Plausibility:1

----------------------------
```

(The MSRANK function is used, with this parameterization for the half-order theory, to rank the structures. On this ranking, structure #3, which scores 76% of the maximum possible score, clearly stands out as providing the best rationalization for the observed spectrum.)

```
|
|    Ranked list of structures:
|    structure#  score
|
|    3              76
|    1              53
|    4              41
|    12             34
|    10             30
|    8              30
|    7              24
|    11             22
|    5              22
|    2              21
|    9              18
|    6              18
|    Do you want to use the EXAMINE option for scored structures?Y
|
```

(The unknown was suspected to be a conjugate of an amino acid and an organic acid. We maintain a small library of protein amino acid skeletons and can use the EXAMINE function to determine which of the twelve generated structures contains any of the known skeletons held in the library.)

```
|
|    Do you want to use a library?Y
|    FILE NAME:(SMITH)AMINOACID.LIBRARY
|    Do you want to enter new selection features?:N
|    The structures have been ranked.
|    MINIMUM       AVERAGE    AND MAXIMUM SCORES.
|    18                32                76
|    ALA present in 1 structures.
|    GLY present in 1 structures.
|    VAL present in 0 structures.
|    LEU present in 0 structures.
|    ILEU present in 0 structures.
|    THRE present in 0 structures.
|    PHE present in 1 structures.
|    TYR present in 0 structures.
|    PRO present in 0 structures.
|    OH-PRO present in 0 structures.
|    ASP present in 0 structures.
|    GLU present in 0 structures.
|    BETA-ALA present in 1 structures.
|    SER present in 0 structures.
|
```

(Once the program has reported on the frequency of occurrence of the standard amino-acid skeletons, the SELECT command can be used to retrieve those structures with particular combinations of features. Here, it is used to select any structures with PHE, GLY, beta-ALA or ALA amino-acid skeletons.)

Example.

```
Enter commands for selecting subsets of structures with
particular features.
->SELECT
>ANY ALA GLY PHE BETA-ALA
4 STRUCTURES WITH (ANY ALA GLY PHE BETA-ALA)
->SCORES
SCORES FOR CURRENT STRUCTURES
#3      76
#4      41
#10     30
#8      30
->DRAW #3

3
```

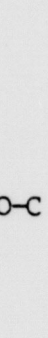


Figure 7-15 shows the structures of the four candidates possessing known amino acid skeletons.

FIGURE 7-15 about here.

Subsequent synthesis confirmed that the top-ranked structure #3, N-acetylphenylalanine methyl ester, was in fact the structure of the unknown. Literature work revealed that this compound has been detected at variable, low levels in the urines of persons suffering from PKU, but has been observed only in small quantities except in cases where the person has been loaded with phenylalanine and never in a patient on a low phenylalanine diet.

This particular problem was small and did not require many of the more powerful features within CONGEN, such as the facilities for specifying constraints on the assembly of superatoms. All the initial data interpretation had to be done by the chemist, e.g. the identification of superatom BZ from the prominent m/z 91 peak in the mass spectrum, but the remaining steps of the identification process

were fully automated. We make no claim that the programs did something that an intelligent chemist could not do. However, the same systematic approach to larger, more complex problems, does, we feel save time and increase the specificity of structural assignment.

CONGEN, and other programs described in this paper, have been made available as a research resource for the general scientific community (50). CONGEN, as implemented on the SUMEX computer facility, may be accessed via ARPANET or TYMNET. The existing version of CONGEN is mainly in the INTERLISP language with specialized subsystems in FORTRAN and in SAIL. This version relies heavily on the large virtual memory and other special features of the TENEX-DEC10 computer system at SUMEX. A production version of the program is currently under development. All components of the new system will be in the BCPL programming language (a simple, widely available, ALGOL-like language); this, together with certain redesign of both algorithms and data-structures, will yield an efficient, compact, and portable version of CONGEN.

F Conclusions.

Complete automation of scientific inference, in the area of structure elucidation based on mass spectral data, is possible for certain, well-characterized classes of chemical compounds. Extension of the fully automated methods to the more general and realistic problem of complex, polyfunctional molecules is unrealistic given current chemical and mass spectral knowledge. A different, only partially automated, approach is necessary for such more complex problems. In this approach, the chemist is an integral part of the problem-solving process contributing to the problem-solving task through his knowledge of chemical constraints and his ability to interpret structural information from a variety of sources.

We reiterate some of the important points made in the paper concerning our methods:

1) The solution of typical structure determination problems requires the use of a variety of information from spectroscopic techniques, chemical treatments and common-sense chemical reasoning. Computational methods are most useful if they can exploit all available information.

2) Interactive computer programs which are under the guidance of a chemist familiar with a particular structural problem can be of valuable assistance in helping identify the structure;

3) We have found it most useful to concentrate our programming efforts on those parts of the inference task which are most difficult to perform manually, e.g., structure generation, determining complete sets of possible fragmentation processes. This frees chemists to devote their time to more creative aspects of problems;

4) Concerning utilization of mass spectral data, we have used the data together with fragmentation rules to rank structural candidates based on the extent of agreement between predicted and observed spectra. This has proven more successful than attempts to develop general computational methods for utilizing the data to obtain structural inferences in the first place.

We do not claim that the interactive programs we have described represent how people actually solve problems. Nor do we claim that computer programs are essential to solve structural problems. However, certain advantages of our methods should be clear. Completeness and non-redundancy are essential in structural chemistry, because unambiguous determination of structure is one of the foundations of chemistry as a science. Our programs help guide chemists to a set of plausible structural candidates, provide assistance in focussing on the correct structure within the set and give a guarantee that no alternatives have been overlooked. Thus, we feel that our methods can help solve some problems more quickly and accurately, thereby increasing chemists' productivity.

G References

19) D.H. Smith, B.G. Buchanan, R.S. Engelmore, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg and C. Djerassi.
J.Am.Chem.Soc.,94,5962 (1972).

20) R.E. Carhart, D. H. Smith, H. Brown and C. Djerassi.
J.Am.Chem.Soc.,97,5755 (1975).

21) D.H. Smith, B.G. Buchanan, W.C. White, E.A. Feigenbaum, C. Djerassi and J. Lederberg.
Tetrahedron,29,3117 (1973).

22) B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg and C. Djerassi.
J.Am.Chem.Soc.,96,6168 (1976).

23) A.Newell.
in "Computer Models of Thought and Language."
R.C. Schank and K.M. Colby (Eds).
W.H.Freeman and Company: San Francisco (1973).

24) R.G. Dromey, B.G. Buchanan, D.H. Smith, J. Lederberg and C. Djerassi.
J.Org.Chem.,40,770 (1975).

25) D.H. Smith and R.E. Carhart.
in "Proceedings of the Symposium on Chemical Applications of High Performance Spectrometry."
M.L. Gross (Ed).
American Chemical Society: Washington (1978).

26) J.G. Nourse.
J.Am.Chem.Soc.,(submitted).

27) J.G. Nourse, R.E. Carhart, D.H. Smith and C. Djerassi.
J.Am.Chem.Soc.,(submitted).

28) A.M. Duffield, A.V. Robertson, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum and J. Lederberg.
J.Am.Chem.Soc.,91,2977 (1969).

29) G. Schroll, A.M. Duffield, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum and J. Lederberg.
J.Am.Chem.Soc.,91,7440 (1969).

30) A. Buchs, A.M. Duffield, G. Schroll, C. Djerassi, A.B. Delfino, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum and J. Lederberg.
J.Am.Chem.Soc.,92,6831 (1970).

References.

31) A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg.
Helv.Chim.Acta,53,1394 (1970).

32) D.H. Smith, B.G. Buchanan, R.S. Engelmore,H. Adlercreutz and C.Djerassi.
J.Am.Chem.Soc.,95,6078 (1973).

33) J. Lederberg.
"DENDRAL-64 A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs."
Technical report to NASA, CR57092 (1964).

34) J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield and C. Djerassi.
J.Am.Chem.Soc.,91,2973 (1969).

35) D.H. Smith.
J.Chem.Inf.Comp.Sci.,15,203 (1975).

36) L.M. Masinter, N.S. Sridharan, R.E. Carhart and D.H. Smith.
J.Am.Chem.Soc.,96,7702 (1974).

37) L.M. Masinter, N.S. Sridharan, R.E. Carhart and D.H. Smith.
J.Am.Chem.Soc.,96,7714 (1974).

38) R.E. Carhart, D.H. Smith, H.Brown and N.S.Sridharan.
J.Chem.Inf.Comp.Sci.,15,124 (1975).

39) R.E. Carhart and D.H. Smith.
Computers in Chemistry,1,79 (1976).

40) R.E. Carhart.
J.Chem.Inf.Comp.Sci.,16,82 (1976).

41) D.H. Smith and R.E. Carhart.
Tetrahedron,32,2513 (1976).

42) C. Cheer, D.H. Smith, C. Djerassi, B. Tursch, J.C. Braekman and D. Daloze.
Tetrahedron,32,1807 (1976).

43) H. Brown, L.M. Masinter and L. Hjelmeland.
Discrete Mathematics,7,1 (1974).

44) H. Brown and L.M. Masinter.
Discrete Mathematics,8,227 (1974).

45) H.Brown.
SIAM J.Appl.Math.,32,534 (1977).

References.

46) B.G. Buchanan, A.M. Duffield and A.V. Robertson.
in "Mass Spectrometry: Techniques and Applications."
G.W.A. Milne (Ed).
John Wiley & Sons. (1971).

47) T.H. Varkony, D.H. Smith and C. Djerassi.
Tetrahedron,34,841 (1978).

48) S. Hammerum and C. Djerassi.
Tetrahedron,31,2391 (1975).

49) L.L. Dunham, C.A. Henrick, D.H. Smith and C. Djerassi.
Org. Mass Spectrom.,11,1120 (1976).

50) R.E. Carhart, S.M. Johnson, D.H. Smith, B.G. Buchanan,
R.G. Dromey and J.Lederberg.
in "Computer Networking and Chemistry."
P.Lykos (Ed).
American Chemical Society: Washington (1976).

7-6) Symbolic representation of the fragmentation rules employed by PLANNER for estrogens.

7-7) The low-resolution mass spectrum of estrone (1). The ions corresponding to the fragmentation rules of Figure 7-6 are marked. The results from PLANNER'S ANALYSIS phase are summarized in Table 7-III.

7-8) The low-resolution mass spectrum of hexanal. Elemental compositions were determined by accurate mass measurements in a subsequent experiment at high resolving power. The elemental composition of the molecular ion and the eight fragment ions shown together formed the input to the MDGGEN program.

7-9) "Mass Distribution Graphs" illustrating various ways of obtaining a C_3H_7 ion from a $C_6H_{12}O$ molecular ion under different assumptions concerning the number of bonds cleaved, number of steps, and numbers of hydrogens transferred in the fragmentation process.

7-10) Side-chain portions of fucosterol (5) and one of its isomers. Bond break plausibilities, differentiating between vinylic and allylic cleavages, have to be defined before the MSRANK program can use the C-22 - C-23 cleavage process to discriminate amongst these isomers.

7-11) The structures of the five macrolide antibiotics whose spectra were analyzed by MSRANK using half-order theory and the class-specific rules of Figure 7-12.

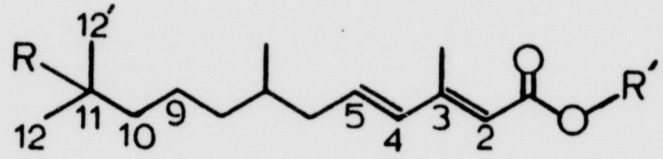
7-12) The class-specific rules used for the analysis of the macrolide antibiotics (structures 6 - 10) involved all pairwise combinations of the illustrated single-bond cleavages.

7-13) The low-resolution mass spectrum of ethyl(2E,4E)-3,7,11-trimethyl-2,4-dodecadienoate (12).

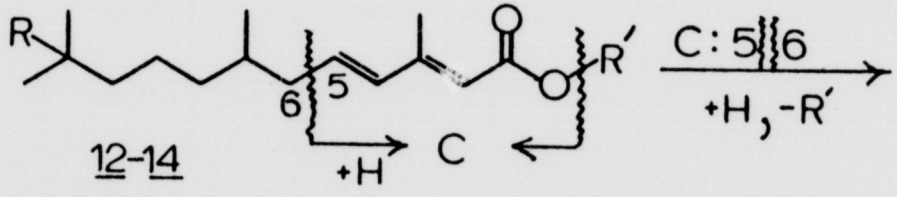
7-14) The low-resolution mass spectrum of n-propyl(2E,4E)-11-methoxy-3,7,11-trimethyl-2,4-dodecadienoate (13).

7-15) The structures and ranking scores of those candidate molecules, for the phenylalanine metabolism problem, possessing known amino acid skeletons.

Index

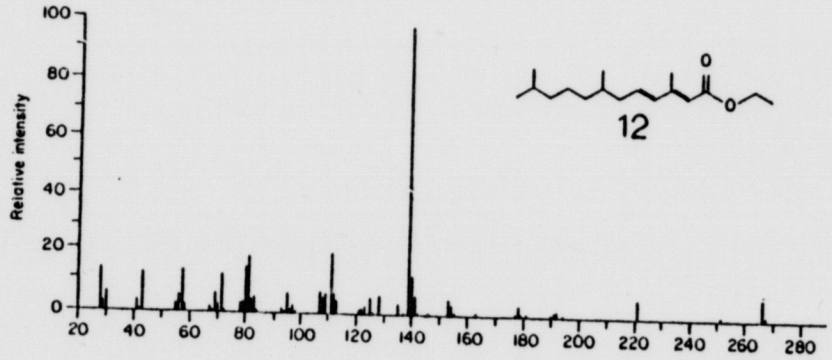


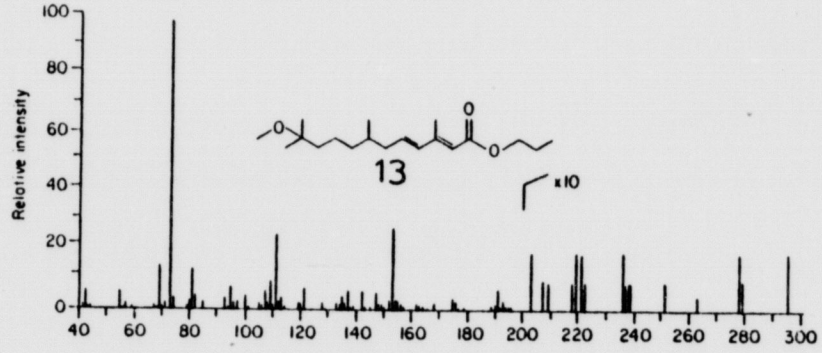
11	R = H; R' = CH_3
12	R = H; R' = C_2H_5
13	R = CH_3O; R' = n-C_3H_7
14	R = CH_3O; R' = $CH(CH_3)_2$

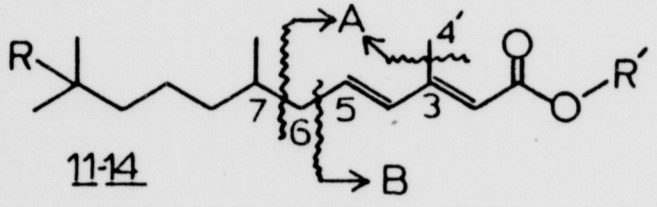


12-14

m/z 111 $[C_6H_7O_2]^{+\cdot}$; ${}_8\Sigma_{40}$ 3.7-4.1


Relative intensity
100
80
60
40
20
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
O
O
12


Relative intensity
100
80
60
40
20
0
40
60
80
100
120
140
160
180
200
220
240
260
280
300
O
O
O
13
x10



11-14

A: 3‖4/6‖7 (+H)

B: 5‖6

Compound	Mass	Composition	$_{8}\Sigma_{40}$
11	125	$C_7H_9O_2$	23.4
12	139	$C_8H_{11}O_2$	19.3
13	153	$C_9H_{13}O_2$	4.8
14	153	$C_9H_{13}O_2$	2.9